VNU Journal of Science:
Computer Science and Communication Engineering
Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

ViMRC - VLSP 2021: Using XLM-RoBERTa and Filter Output for Vietnamese Machine Reading Comprehension

Dang Van Nhan*, Nguyen Le Minh

University of Information Technology, Vietnam National University, Ho Chi Minh City
Quarter 6, Linh Trung, Thu Duc, Ho Chi Minh City, Vietnam

Received 28 December 2021
Revised 31 March 2022; Accepted 5 May 2022

Abstract: Nowadays, the amount of information has become huge, and our task is to find the correct answers to the questions. In fact, not every question has an answer, and then the best answer should be don't know, where the model that makes the prediction is the empty string. Building a high-accuracy response model will make people's lives easier. We have the SQuAD dataset for English that helps train the machine reading comprehension model. Based on SQuAD 2.0, the organizing committee developed the Vietnamese Question Answering Dataset UIT-ViQuAD 2.0 [1], a reading comprehension dataset consisting of questions posed by crowd-workers on a set of Wikipedia Vietnamese articles. The UIT-ViQuAD 2.0 dataset evolved from version 1.0 with the difference that version 2.0 contained answerable and unanswerable questions. The challenge of this problem [2] is to distinguish between answerable and unanswerable questions. The answer to every question is a span of text from the corresponding reading passage, or the question might be unanswerable. Our system employs simple yet highly effective methods. The system uses a pre-trained language model (PLM) called XLM-RoBERTa (XLM-R [3]), combined with filtering results from multiple output files to produce the final result. We created about 5-7 output files and selected the most repetitions as the final prediction answer. After filtering, our system increased from 75.172% to 76.386% at the F1 measure and achieved 65,329% in the EM measure on the Private Test set,…

Keywords: Machine-reading-comprehension, VSLP, MRC, Vietnamese.

1. Introduction

Finding the correct answer is a daily need of every person. Therefore, building a high-precision answering system plays a vital role in human life. The MRC is an integral part of the open-domain question answering system, and the accuracy of the answer depends on this

* Corresponding author.
E-mail address: 18521172@gm.uit.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.336

problem. We believe that MRC is a critical area for natural language processing (NLP) and greatly influences other fields in NLP. The task has achieved outstanding results by introducing quality MRC datasets such as SQuAD 1.0 [4], CMRC [5], NewsQA [6], UIT-ViQuAD 1.0 [7]. Many systems have even surpassed human performance. However, the machine reading comprehension models that deal with answerable and unanswerable questions in Vietnamese are currently quite a few. Therefore, we aim to build a system to predict answers with high accuracy based on the UIT-ViQuAD 2.0 dataset provided by the organizers of the VLSP-ViMRC 2021. Table 1 gives an example of an answerable and unanswerable question in the UIT-ViQuAD 2.0 dataset.

The rest of the paper is structured as follows. Section 2 introduces the related work that we researched for the construction of the MRC system. Section 3 talks about our system and how we build it. Next are the parameters of the UIT-ViQuAD 2.0 dataset, and our analysis of the dataset is presented in section 4. Section 5 presents the experimental results and the measure of the problem in the competition. Finally, in section 6, we conclude and talk about the development directions for our system.

2. Related Work

Machine Reading Comprehension is a necessary problem in today's life and appeared a long time ago. However, because previous datasets had not achieved high accuracy, there was a time when the problem settled down. Until 2015, when the SQuAD (Wikipedia-sourced Machine Reading Comprehension Dataset) dataset was released, the models achieved high accuracy thanks to the properly constructed dataset. Following that success, datasets for many other languages were also gradually born. The Vietnamese machine reading comprehension dataset has only appeared in the last three years. Prominent Vietnamese MRC datasets are:

UIT-ViNewsQA: Vietnamese Corpus for Machine Reading Comprehension of Health News Articles published in January 2020.

UIT-ViQuAD: The first dataset sourced from the Vietnamese Wikipedia was published in September 2020. The accuracy for the problem in version 1.0 of UIT-ViQuAD currently achieves the highest accuracy of 89.54%.

UIT-ViMMRC: dataset for Vietnamese reading comprehension by answering multiple-choice questions published in October 2020.

UIT-ViWikiQA: is converted from the UIT-ViQuAD dataset for evaluating sentence extraction-based machine reading comprehension in the Vietnamese language. The dataset released in May 2021.

In this paper, we work with the UIT-ViQuAD 2.0 dataset. Compared to the above datasets, UIT-ViQuAD 2.0 includes unanswerable questions, similar to SquAD 2.0 dataset. However, version 2.0 causes more difficulties for the model, as the model needs to correctly distinguish the cases where the question is answerable or not. The wrong discrimination greatly influences the accuracy of the model because, for the wrong prediction sentence, the accuracy of that sentence is 0%.

We need to solve two big problems for the Vietnamese machine reading comprehension problem. The first is machine reading comprehension. We need to build a system that can determine if a question is answerable or not. If it can be answered, extract the answer span from the passage. The second is the complexity of the Vietnamese language. The model needs to be trained on a large amount of Vietnamese data to be able to predict the answer with high accuracy. A more straightforward way is to use a PLM. Transfer learning helps us to inherit pre-trained parameters, saving time while ensuring performance.

Transfer learning is to transfer the learned features of the previous neuron to the following neurons without re-learning. It is similar to a teacher 'transfer' on her knowledge to the students. A pre-trained model is a saved network

that has been previously trained on a large data set. Bidirectional encoders in pre-trained model can be used to generate contextualized representations of input embeddings using the entire input context, can be learned using a masked language model objective where a model is trained to guess the missing information from an input. Pretrained language models can be fine-tuned for specific cases, each application can be added different lightweight classifier layers on top of the outputs of the pretrained model. With the development of transfer learning in the natural language processing (NLP) field, pre-trained models are preferred because it saves training time on large amounts of data to the model can handle a particular language and task well. We have studied some popular models giving good results on Vietnamese topics, MRC topics in general, and Vietnamese MRC in particular, for example:

BARTpho [8]: is the homonym of the word "bowl of Pho"in Vietnamese, using the "large"architecture and pre-training scheme of the seq-to-seq denoising model BART, which is specifically suitable for this NLP tasks. Experiments on a downstream task of Vietnamese text summarization show that BARTpho outperforms the strong baseline mBART and assesses the state-of-the-art in both automated and human evaluations.

PhoBERT [9]: are the state-of-the-art language model for Vietnamese ("Pho"is a popular food in Vietnam). Test results show that PhoBERT gives good results in many Vietnamese-specific NLP tasks, including Part of Speech tagging, Dependency Parsing, Named Entity Recognition, and Natural Language Inference course.

Table 1. Answerable and unanswerable example in UIT-ViQuAD 2.0

Passage: Mã máy nhị phân (khác với mã hợp ngữ) có thể được xem như là phương thức biểu diễn thấp nhất của một chương trình đã biên dịch hay hợp dịch, hay là ngôn ngữ lập trình nguyên thủy phụ thuộcvào phần cứng (ngôn ngữ lập trình thế hệ đầu tiên). Mặc dù chúng ta hoàn toàn có thể viết chương trìnhtrực tiếp bằng mã nhị phân, việc này rất khó khăn và dễ gây ra những lỗi nghiêm trọng vì ta cần phải quản lý từng bit đơn lẻ và tính toán các địa chỉ và hằng số học một cách thủ công. Do đó, ngoại trừ những thao tác cần tối ưu và gỡ lỗi chuyên biệt, chúng ta rất hiếm khi làm điều này. **English**: Binary machine code (different from the assembly code) can be viewed as the lowest representation of a compiled or compatible program, or the primitive programming language depends on the hardware (first-generation language programming). Although we can completely write programsdirectly with binary code, this is very difficult and easy to cause severe errors because we need to manage every bit and calculate addresses and manually learned constants. Therefore, except for optimal and debugging operations, we rarely do this.
Question 1: "Ngôn ngữ lập trình thế hệ đầu tiên là ngôn ngữ gì?" **English**: "What is a first-generation programming language?" **Answerable**: "Mã máy nhị phân" **English**: "Binary machine code"
Question 2: "Ngôn ngữ lập trình hợp ngữ đầu tiên là ngôn ngữ gì?" **English**:"What was the first assembly language?" Unanswerable: "" **Plausible answer**: "Mã máy nhị phân" **English**: "Binary machine code"

XLM-R model [10]: proposed in Unsupervised Cross-lingual Representation Learning at Scale by Alexis Conneau et al. XLM-R is primarily based on Facebook's RoBERTa model [11], released in 2019. It is a large multilingual language model trained on 2.5TB of filtered CommonCrawl data and is the XLM [12] model's state-of-the-art. XLM-R shows the ability to train multiple language models (including Vietnamese) without

sacrificing per-language performance. Models such as XLM and mBERT are limited in learning valuable representations for low-resource languages. XLM-R improves upon previous multilingual approaches by combining lots of data and training over 100 languages - including so-called low-resource languages, which lack tagged datasets labeled and are not widely labeled. Unlike some XLM multilingual models, it now no longer requires the language controller to recognize which language is used and decide the correct language from the input id.

We decided to use two pre-trained models, PhoBERT and XLM-R for our experiments. Experimental results show that XLM-R gives better results in this Vietnamese MRC task, so we used XLM-R.

3. Our System

3.1. Model Based on XLM-R

Our system Figure 1 is built with the backbone of the pre-trained language model (PrLM) XLM-R Figure 2.

XLM-R is a model trained in about 100 languages, including Vietnamese. Therefore, we also download the model and go through the same processing steps as the SQuAD 2.0 dataset, including preprocessing, tokenizing, feature extraction, and training. However, because Vietnamese is not the same as English, hyper-parameters such as epoch number, batch size, or learning rate will differ. Therefore, we need to test many times to choose the most suitable parameters.

First, we load the data from the train set into clusters: Context C = {C_1, . . . ,C_n}, question Q = {Q_1, . . . , Q_n} and answer A = {A_1, . . . , A_n}. The test dataset with input includes the clusters: context and question. Our task is to train the model on the training dataset so that the model can find the correct answer to the question (the answer is null text or a span extracted from the paragraph). For the model to be able to understand human language, we need to convert the passages, questions, and answers from text to numbers. We use a tokenizer language model compatible with a pre-trained language model from XLM-R. Then, we proceed to process the data to extract the necessary features for the model with pre-written functions from the Hugging face for the SQuAD 2.0 dataset because the structure of the UIT-ViQuAD 2.0 dataset is the same as the SQuAD 2.0 dataset. The model we use to train the MRC task is RoBERTa because it is a model that can take advantage of the parameters from PrLM XLM-R. We feed the features extracted from the training dataset for model and train. The model after being trained to determine if the question has an answer within the passage. For answerable questions, the resulting span extracted from the context is calculated by taking the answer's start position and end position values with the highest probability among the positions that the model predicts. Cases where the model returns null text, include: The answer is within the question. The answer has a larger starting position than the ending position and a probability less than the null threshold.

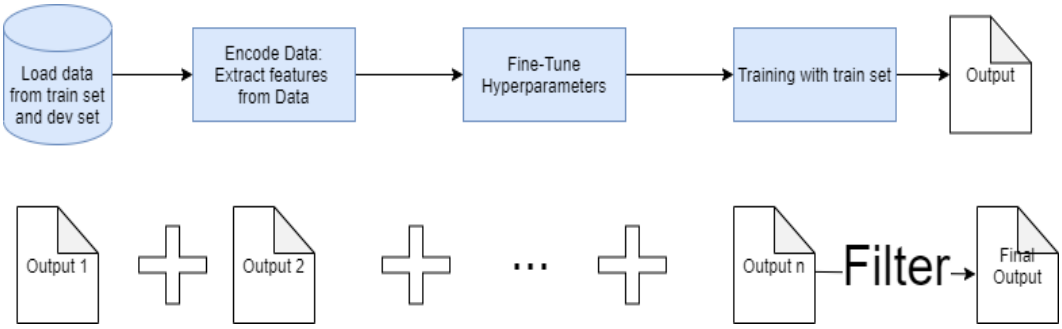

Figure 1. Our Vietnamese Machine Reading Comprehension model.

Figure 2 shows how the XLM-R pre-trained language model takes input and gives an answer. The model will receive the question and the context (document), tokenize them and separate the question and the context with [SEP]. After processing and calculating, the model finds the starting and ending positions of the answer. If the model predicts the question has no answer, it returns null text; otherwise, it will return multiple start and end positions. The selected answer is a span in the context with the start position value multiplied by the end position with the highest result.

3.2. Filter Module

The difference of our system from existing models using XLM-R is that we built an additional module to process and calculate consensus for the answers. When the model undergoes many different training times, the results at each prediction time are also different. When going through labeling rounds, humans also give the same answers on easy questions, and for complex questions, maybe each person will answer a different answer. At that time, people tend to take the outcome as the most similar answer among the predictors. This approach does not always find the correct answer, but the correct rate will be increased in general. As a result, the model can eliminate less likely answers. This module is applicable not only to the XLM-R model but also to all other models since the method is based on the highest probability, helping to maximize the answer.

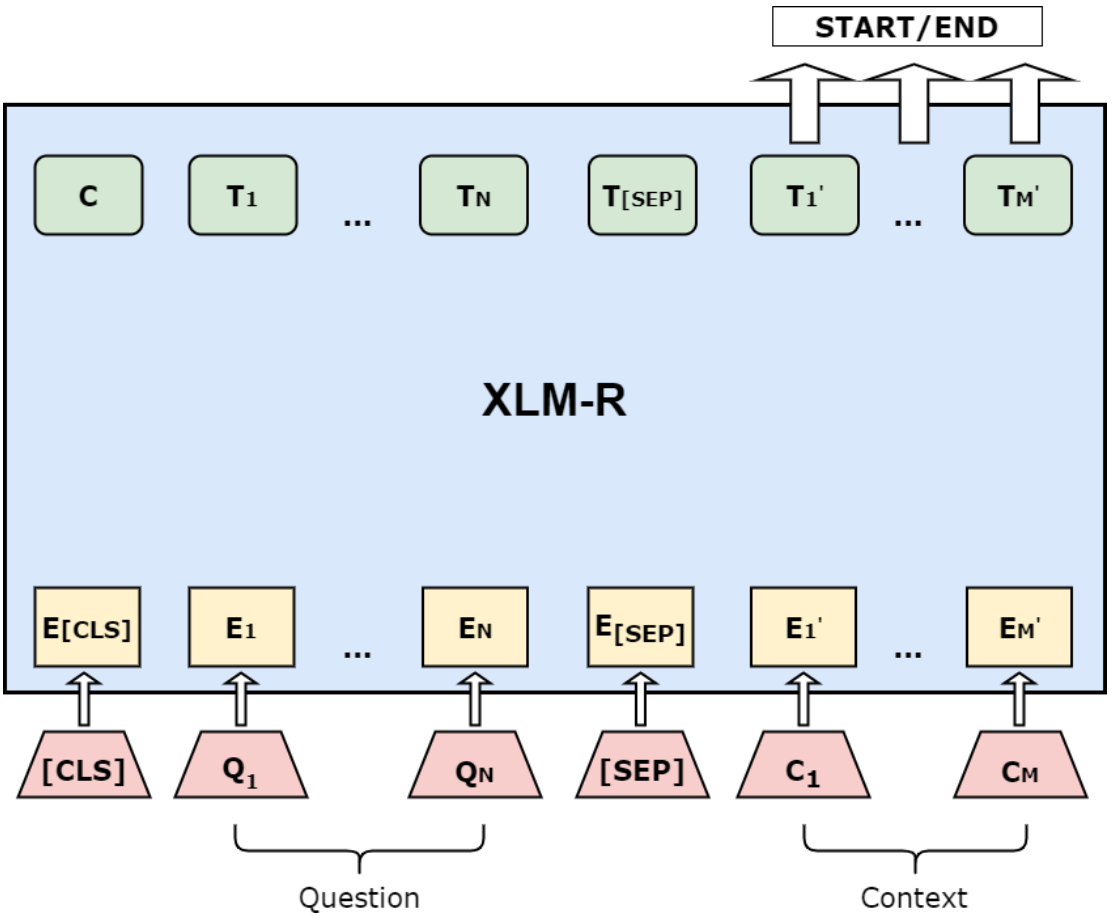

Figure 2. The XML-R Reading Comprehension model.

This module is inspired by creating a dataset: an answer in data must be agreed from at least two annotators. Criteria for selecting output files are files that do not overlap. At each model training, there should be differences in hyperparameters. We found that the prediction files achieved the highest accuracy at three epochs or four epochs during the experiment. So, at each guess, we save these two prediction result files. Every time we train, we change the hyperparameters like learning_rate, batch_size, max_seq_length, max_query_length. After three runs of the model, we get six output files:

$f1 = \{A_{11}, \ldots A_{1n}\}$, ... , $f6 = \{A_{61}, \ldots, A_{6n}\}$

In which: $f1, \ldots f6$ are prediction files. A is the predicted answer corresponding to each question in the file. Our final output file after filtering is:

$$f = \{A_1, \ldots A_n\}$$

In which:

A_i = max_repeat_answer $\{A_{1i}, \ldots, A_{6i}\}$ $(1 \leq i \leq n)$ According to our observations in the prediction files, the answers to difficult questions often vary, but repeated answers in multiple files are usually the correct answer. Therefore, we chose the answer in the last file as the most repeated answer from six files. We randomly selected the final answer for the case where all six files predict six different answers for the same question. Experiments show that when taking the answers with the most frequency, our model increases from 1-2% accuracy on the test dataset of UIT-ViQuAD 2.0.

Table 6 illustrates how the filter module selects the final result from the 4 answers. The answer "sông ba˘ng c.c nam cua Hoa Ky`"" was chosen because it repeats the most times.

4. Dataset

4.1. Statistics

The initial version of UIT-ViQuAD 2.0 was UIT-ViQuAD 1.0, a dataset developed based on SQ 1.0. The dataset contains 23 K+ question-answer pairs on 170+ articles extracted from Wikipedia. The titles from the dataset are taken from high-ranking Wikipedia articles; each title is divided into several paragraphs, each paragraph has many questions. A continuous span extracted from the passage corresponds to each question with an answer.

The dataset used in this paper is from the Vietnamese Machine Reading Comprehension task at The 8th International Workshop on Vietnamese Language and Speech Processing (VLSP 2021) created by [7].

UIT-ViQuAD 2.0 combines the 23K questions in UIT-ViQuAD 1.0 with over 12K unanswerable questions written adversarially by crowd-workers to look similar to answerable ones. The difference from version 1.0 is that each question can answer or not. Similar to the set of UIT-ViQuAD 1.0, the answer to each question is a span extracted from the passage. We consider the answer as a null text for the unanswerable question and add a plausible answer field containing the confusing answer that the model might choose. With the UIT-ViQuAD 2.0 dataset, MRC systems must answer questions when possible and determine when the context supports no answer.

Table 2 shows that the total number of questions in the UIT-ViQuAD 2.0 is 35.990. In addition, the table also lists the number of articles, passages, and unanswerable questions in the train, dev, and test set of the dataset.

The external dataset UIT-ViQuAD 2.0 helps to train the model to find the answer to the question, as well as the normal person, there will be cases where there are questions that cannot be answered. The datasets also have such a case section that helps build more complete models in many ways, not just finding the right answer.

4.2. Vietnamese Characteristics

Vietnamese is a complex language in many aspects, such as tones (Vietnamese has six tones), reading, and context. Some examples:

- ca, cá, cà, cạ, cả, cã each word has a different meaning.
- nghỉ, nghĩ have the same pronunciation.

- y and i have the same reading, but when in the word, the word, the reading and meaning are different.
- cổ means both old, neck, her depending on the context.

Each of the above factors changes the meaning of Vietnamese, so if the data is not processed, it will significantly affect the resolution. Therefore, we have to check the Vietnamese dataset to make sure it doesn't have too many problems affecting problem-solving.

Question words in Vietnamese are more diverse than in English, and Vietnamese also has questions: What, Where, When, Who, Which, How. With each question word in Vietnamese, there are more ways to ask. For example, for the What question, in Vietnamese, there are related questions such as gì, là gì, cái gì, để làm gì. Same for Where, When, Who, Which, How. Example "What is this?", when translated into Vietnamese, there are many ways to ask, such as "Cái gì đây?", "Đây là con gì?", "Đây là gì?". Therefore, the model is more challenging to understand the question in Vietnamese.

Comma position plays a significant role in Vietnamese. Just changing the meaning of a sentence can be changed entirely. E.g.:

- "Rắn hổ mang bò lên núi".
- "Rắn, hổ mang bò lên núi."

When changing the commas in the two sentences above, the sentence's meaning changescompletely. In Vietnamese, a sentence's meaning depends on the surrounding context. Because a sentence can have many meanings. E.g.: "Cổ đeo vòng cổ."

The above sentence means "a girl is wearing a necklace around her neck"or "a girl is wearing her necklace"or "a girl is wearing an antique necklace". The above aspects make Vietnamese more complicated for the model. Especially with questions using synonyms, homonyms, or equivalent knowledge, the model will be ambiguous when learning.

5. Results and Evaluation

5.1. Evaluation Metrics

Similar to the evaluation method on the SQuAD 2.0 dataset, UIT-ViQuAD 2.0 also uses EM and F1-Scores as assessment measures for the Vietnamese machine reading comprehension task.

Table 2. Overview statistics of the UIT-ViQuAD 2.0 dataset

	Train	Public Test	Private Test	All
Number of articles	138	19	19	176
Number of passages	4,101	557	515	5,173
Number of total questions	28,457	3,821	3,712	35,990
Number of unanswerable questions	9,217	1,168	1,116	11,501

Table 3. Detailed results on the private dev, evaluated on EM and F1 scores

F1	EM	Has Answer F1	Has Answer EM	No Answer F1	No Answer EM
70,764	56,223	65,724	44,329	81,457	81,457

F1-Score: F1-score is a popular metric for natural language processing and is also used in machine reading comprehension. F1-Score is a more objective representation of the performance of a model. F1-score estimated over the individual tokens in the predicted answer against those in the gold standard answers. The F1-score is based on the number of matched tokens between the predicted and gold standard answers. To calculate the F1 measure, we treat

each gold standard answer and predicted answer as a bag of tokens. Then calculate the number of tokens that are the same (numSame) between predicted answers and gold standard answers.

Recall: calculated by scaling the same tokens by the tokens in the gold standard answer.

$$recall = \frac{numSame}{numTruth} \quad (1)$$

Precision: calculated by scaling the same tokens by the tokens in the predicted answer.

$$precision = \frac{numSame}{numPredict} \quad (2)$$

The formula calculates the measure F1:

$$f_1 = \frac{2 * recall * precision}{(recall + precision)} \quad (3)$$

Exact-match: Each answer where the prediction tokens perfectly match the token of the gold standard answer will get a value of 1, otherwise get a value of 0. Divide all correct answers by the total number of questions in the test dataset, and we get the EM measure result. Note, before comparison, the text of the predicted answer and the gold standard answer is normalized.

In VLSP 2021 - the task of Vietnamese MRC, the final ranking is evaluated on the test set, according to the F1-score (EM as a secondary metric when there is a tie). The results are round to the nearest hundredth (3 decimal places). If two teams have the same F1 score, EM score is used to determine which team is better.

Table 4. Aggregate results on the Public Test, Private Dev and Private Test, evaluated on EM and F1 scores

	Public Test		**Private Dev**		**Private Test**	
	F1	**EM**	**F1**	**EM**	**F1**	**EM**
PhoBert	69,280	58,493	63,130	49,363	-	-
XLM-R	**78,637**	**68,804**	70,764	56,223	75,172	63,147
XLM-R with output filter	-	-	-	-	**76,386**	**65,329**

5.2. Result

We performed on the public test dataset and the private test dataset of the contest and got the result of 4th overall on the private test dataset the private test dataset of the contest and got the result of 4th overall on the private test dataset (Table 5).

We have not applied the output file filtering in the public test phase. After using the output file filtering, it can be seen that our team's ranking is significantly improved on the leaderboard.

We found that PrLMs play a vital role in predicting the results for the Vietnamese MRC problem through the experimental process. According to the PrLMs the groups provided to the organizers of the VLSP2021-MRC and announced by the organizers, most of the groups use the currently prominent PrLMs suitable for Vietnamese for the MRC task such as XLM-R, PhoBERT, BARTPho, mBERT, and mT5 [13]. The fact from the SQuAD 2.0 dataset shows that the leading models are often combined from many different models. Our system, currently, is only using a single model in combination with the filter output module.

5.3. Analysis

We decided to test two models that we think are the most feasible for the Vietnamese language: PhoBERT and XLM-RoBERTa. Our experimental results are shown in table 4 on three sets, of which two sets from the contest are Public Test and Private Test. In the Private Dev set, we split ourselves to evaluate the model. Experimental results with PhoBERT and XLM-RoBERTa on the Public test set and Private Dev set show that XLM-RoBERTa has superior results. Therefore, we decided to use the XLM-

RoBERTa model for the final test set of the VLSP-MRC 2021. With the addition of the filter module, we significantly improved our rankings in the competition, which is also our best result in this contest. Table 3 experiments on the Private Dev set that we divided ourselves showed that the F1 and EM measures had a relatively significant difference: 70.764% and 56.223%. The reason is that the answers with the correct answer are predicted with low accuracy. As we guess, the model predicted many unanswerable questions, while gold data are answerable. Such cases will receive 0% of the EM measure, resulting in the answer being shallow: 44.329%. Therefore, to increase the model's accuracy, the model must solve these cases.

Table 5. Ranking of participating teams on public test set and private test set

Public test			**Private test**		
Team	**F1**	**EM**	**Team**	**F1**	**EM**
NLP_HUST	**84,236**	**77,728**	**vc-tus**	**77,241**	**66,137**
NTQ	84,089	77,990	ebisu_uit	77,222	67,430
ebisu_uit	82,622	73,698	F-NLP	76,456	64,655
vc-tus	81,013	71,316	**UIT-MegaPikachu**	**76,386**	**65,329**
F-NLP	80,578	70,662	SDSOM	75,981	63,012
SDSOM	79,594	69,092	UITSunWind	75,587	64,871
UITSunWind	79,130	69,720	Big Heroes	74,241	61,126
UIT-MegaPikachu	**78,637**	**68,804**	914-clover	73,027	61,853
914-Clover	78,515	69,013	NTQ	72,863	60,938
Big Heroes	78,491	68,150	Hey VinMart	70,352	57,786
BASELINE	**63,031**	**53,546**	**BASELINE**	**60,338**	**49,353**

Table 6. Example how the filter module chooses the final answer

Context: Dãy núi Sierra Nevada (tức "dãy núi tuyết"trong tiếng Tây Ban Nha) ở phía đông và trung tâm tiểu bang, có núi Whitney là đỉnh núi cao nhất trong 48 tiểu bang (4,421 mét (14,505 feet)). Trong dãy Sierra còn có Công viên Quốc gia Yosemite và hồ Tahoe (một hồ nước ngọt sâu và là hồ lớn nhất của tiểubang theo thể tích). Bên phía đông của dãy Sierra là thung lũng Owens và hồ Mono – nơi sinh sống chủ yếu của chim biển. Còn bên phía tây là hồ Clear, hồ nước ngọt lớn nhất của California theo diện tích.Vào mùa đông, nhiệt độ ở dãy Sierra Nevada xuống tới nhiệt độ đóng băng và ở đây có hàng chục dòngsông băng nhỏ, trong đó có sông băng cực nam của Hoa Kỳ, sông băng Palisade. **English**: The Sierra Nevada (or "snow mountain range"in Spanish) ranges in the east and central parts ofthe state, with Mount Whitney being the tallest peak in the 48 states (4,421 meters (14,505 feet)). In the Sierra Range are also Yosemite National Park and Lake Tahoe (a deep freshwater lake and the state'slargest lake by volume). To the east of the Sierra are the Owens Valley and Mono Lake, which is mainly inhabited by seabirds. To the west is Clear Lake, California's largest freshwater lake by area. In winter,temperatures in the Sierra Nevada drop to freezing, and there are dozens of small glaciers, including the southernmost glacier in the United States, the Palisade Glacier.
Question: "Hãy cho biết vị trí địa lý của sông băng Palisade trong lãnh thổ nước Mỹ?" **English**: "What is the geographical location of the Palisade Glacier in the United States?"
Predict 1: "sông băng cực nam của Hoa Kỳ" **English**:"America's southernmost glacier" **Predict 2**: "dãy Sierra Nevada" **English**:"Sierra Nevada range"

Predict 3: "cực nam"
English:"southernmost" **Predict 4**: "sông băng cực nam của Hoa Kỳ" **English**:"America's southernmost glacier"
Answer after filter: "sông băng cực nam của Hoa Kỳ" **English**: "America's southernmost glacier"

Table 7. The case is no answer but received an answer

Context: "Năm 1996, Hàn Quốc trở thành thành viên của OECD, một mốc quan trọng trong lịch sử phát triển của đất nước. Giống như các quốc gia phát triển khác, ngành dịch vụ đã tăng nhanh, chiếm khoảng 70% GDP. Cùng với sự phát triển về kinh tế, đời sống của nhân dân được nâng cao rất nhanh trở nên ngang bằng các quốc gia phát triển khác ở châu Âu và các nước Bắc Mỹ. Chỉ số phát triển con người (HDI) đạt 0,912 vào năm 2006. Hiện nay, thu nhập và tài sản của Hàn Quốc đang tăng một phầnlà do sự đầu tư và xuất khẩu công nghệ cao sang các nước đang phát triển như Trung Quốc, Việt Nam,và Indonesia". **English**: "In 1996, Korea became a member of the OECD, a milestone in the country's developmenthistory. Like other developed countries, the service industry has grown rapidly, accounting for about 70% of GDP. Along with economic development, people's living standards have been improved veryquickly, becoming comparable to other developed countries in Europe and North America. The Human Development Index (HDI) reached 0.912 in 2006. Currently, Korea's income and wealth are increasing in part due to high-tech investment and exports to developing countries like China, Vietnam, and Indonesia."
Question: "Từ khi nào Hàn Quốc thu nhập thành viên của OECD?" **English**: "Since when does Korea income member of the OECD?"
Wrong Answer: "Năm 1996," **English**: "In 1996,"
Correct Answer: "" **English**: ""

Table 8. The case has an answer but received no answer

Context: "Khi Tướng Park Chung-hee nắm quyền vào năm 1961, Hàn Quốc đã có một thu nhập bình quân đầu người ít hơn $ 80 USD mỗi năm. Trong thời gian đó, Hàn Quốc chủ yếu là phụ thuộc vào viện trợ nước ngoài, chủ yếu là từ Mỹ để đổi lấy sự tham gia của Hàn Quốc trong chiến tranh Việt Nam. Hàn Quốc đã cử khoảng 320.000 quân nhân sang tham chiến cùng Mỹ trong chiến tranh Việt Nam để đổi lấy những khoản viện trợ của Mỹ. Khoảng 5.000 lính Hàn Quốc đã chết và khoảng 11.000 lính khác bị thương tật trong cuộc chiến này. Đội quân này cũng gây ra một danh sáchdài những tội ác chiến tranh, những vụ thảm sát thường dân Việt Nam khi tham chiến,... **English**: When General Park Chung-hee took power in 1961, Korea had a per capita income of less than $80 USD per year. During that time, Korea was largely dependent on foreign aid, mainly from theUnited States, in exchange for Korea's participation in the Vietnam War. South Korea sent about 320,000 troops to fight with the US in the Vietnam War in exchange for US aid. About 5,000 South Korean soldiers died and about 11,000 more were injured in this war. This army also committed a long list of war crimes, massacres of Vietnamese civilians during the war,...
Question: "Tổng thống Park Chung-hee lên nhận chức vào thời gian nào?" **English**: "When did President Park Chung-hee take office?"
Wrong Answer: "" **English**: ""
Correct Answer: "năm 1961" **English**: "In 1961"

Table 9. The case has an answer but received the wrong answer:

Context: "Hàn Quốc cũng là một nước phát triển có sự tăng trưởng kinh tế nhanh nhất, với tốc độ tăng trưởng GDP bình quân là 5% mỗi năm - một phân tích gần đây nhất bởi Goldman Sachs năm 2007 đã chỉ ra nếu duy trì được tốc độ tăng trưởng này, Hàn Quốc sẽ trở thành nước có nềnkinh tế lớn thứ 9 trên thế giới vào năm 2025 với GDP bình quân đầu người là 52.000 USD và tiếp 25 năm sau nữa sẽ vượt qua tất cả các nước ngoại trừ Hoa Kỳ để trở thành nước có GDPđầu người thứ hai trên thế giới, với bình quân đầu người là 81.000 USD." **English**: South Korea is also a developed country with the fastest economic growth, with an average GDP growth rate of 5% per year - a most recent analysis by Goldman Sachs in 2007 showed that if the rate can be maintained at this rate of growth, Korea will become the 9th largest economy in the world by 2025 with a GDP per capita of 52,000 USD and in the next 25 years it will surpass all countries except the United States to become the country with the second GDP per capita in the world, with a per capita of 81,000 USD.
Question: "Theo như phân tích gần đây nhất của Goldman, nếu như tiếp tục tăng trưởng ổn định, nền kinh tế của Hàn Quốc sẽ phát triển như thế nào vào năm 2050?" **English**: "According to Goldman's most recent analysis, if it continues to grow steadily, how willthe Korean economy develop in 2050?"
Wrong Answer: "trở thành nước có nền kinh tế lớn thứ 9 trên thế giới" **English**: "become the country has the 9th largest economy in the world"
Correct Answer: "sẽ vượt qua tất cả các nước ngoại trừ Hoa Kỳ để trở thành nước có GDP đầu người thứ hai trên thế giới" **English**:"will surpass all countries except the United States to become the country with the second GDPper capita in the world"

From our results, we find that there are three error cases, which are:

The first error case is that there is no answer, but the result is that there is an answer. The wrong answers received often answer the right type of question (khi nào (when), cái gì (what),...), but the model cannot distinguish similar words and was confused because it has many of the same words (example in Table 7: thu thập thành viên and trở thành thành viên), leading to misunderstanding the meaning of the sentence.

The next error is having an answer but getting no answer. We found that the reason for this error was that the model did not understand synonyms or equivalent knowledge in the context (example in Table 8: The General and the President here are the same person, Park Chung-hee).

The last error case is an answer, but the result is a wrong answer. We find that this error is that the question requires inference such as computation (example in Table 9: the model cannot calculate that 2050 is 25 years after 2025 so it will confuse it with 2025.) to give the correct answer, which leads to the model answering a given question similar to a question without inference.

In addition, to increase the accuracy of the model. We can apply the ensemble model to combine multiple models, helping to improve performance. Or use more data sets other than UIT-ViQuAD 2.0 to train a more diverse learning model.

6. Conclusion and Future Work

The MRC problem with the UIT-ViQuA 2.0 dataset presents us with many challenges. The first challenge is to distinguish whether a question can be answered or not; determining right or wrong significantly affects accuracy. The second challenge is to find the exact answers to the questions. Questions that use synonyms or equivalent knowledge cause many difficulties for predictive models. Therefore, we need to solve these problems if we want to increase the accuracy of the problem.

In this paper, we have successfully built an answer prediction model based on XLM-RoBERTa. We also contribute a new idea to the machine reading comprehension problem using the filter module. With this model, at the VLSP2021 - the task of Vietnamese MRC

competition, our team (UIT-MegaPikachu) achieved 4th place overall. With a method and idea that is not too complicated but highly effective, the difference is less than 1% in the F1 measure compared to the leading group. This model can be considered as one of the best models to date in the UIT-ViQuAD 2.0 dataset.

However, our model is a single model. We have not applied ensemble for this contest. The use of the ensemble model significantly increases the prediction rate for machine-reading systems, and the current high-ranking models all use ensemble. Therefore, our future goal is to apply the ensemble model and use other modern methods for our system to achieve an accuracy of over 80% for this problem. We then combine it with information retrieval to build a complete open-source question-answering system. Because the issue of machine reading comprehension requires an input passage before asking a question to develop into a product, the applicability is not high. Instead, a model that takes in the question then retrieves the document itself and predicts the answer will be more applicable. With nearly 100 million people, the Machine reading comprehension problem is highly applicable to Vietnamese. The problem can support building a chatbot system or developing into a question-answering application because the human need to find answers happens every day and every hour.

References

[1] K. V. Nguyen, S. Q. Tran, L. T. Nguyen, T. V. Huynh, S. T. Luu, N. L. T. Nguyen, VLSP 2021 – Vimrc Challenge: Vietnamese Machine Reading Comprehension, in: Proceedings of the 8th International Workshop on Vietnamese Language and Speech Processing (VLSP 2021), 2021.

[2] N. V. Kiet, T. Q. Son, N. T. Luan, H. V. Tin, L. T. Son, N. L. T. Ngan, VLSP 2021- ViMRC Challenge: Vietnamese Machine Reading Comprehension, VSLP 2021.

[3] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, 2020, ArXiv:1911.02116.

[4] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ Questions for Machine Comprehension of Text, Arxiv Preprint Arxiv:1606.05250.

[5] Y. Cui, T. Liu, W. Che, L. Xiao, Z. Chen, W. Ma, S. Wang, G. Hu, A Span-Extraction Dataset for Chinese Machine Reading Comprehension, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), http://dx.doi.org/10.18653/v1/D19-1600

[6] A. Trischler, Z. Ye, X. Yuan, J. He, P. Bachman, A Parallel-Hierarchical Model for Machine Comprehension on Sparse Data, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 432-441, https://doi.org/10.18653/v1/P16-1041.

[7] K. Nguyen, V. Nguyen, A. Nguyen, N. Nguyen, A Vietnamese Dataset for Evaluating Machine Reading Comprehension, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 2595–2605.

[8] N. L. Tran, D. M. Le, D. Q. Nguyen, BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese, ArXiv:2109.09701.

[9] D. Q. Nguyen, A. T. Nguyen, Phobert: Pre-Trained Language Models f or Vietnamese, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1037-1042, https://doi.org/10.18653/v1/2020.findingsemnlp.92.

[10] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-Lingual Representation Learning at Scale, Arxiv Preprint Arxiv:1911.02116.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A Robustly Optimized Bert Pretraining Approach, Arxiv Preprint Arxiv:1907.11692.

[12] G. Lample, A. Conneau, Cross-Lingual Language Model Pretraining, Arxiv Preprint Arxiv:1901.07291.